How Do Data Skills Affect Firm Productivity: Evidence from Process-driven vs. Innovation-driven Practices

Lynn Wu
wulynn@wharton.upenn.edu

Lorin Hitt
lhitt@wharton.upenn.edu

University of Pennsylvania
The Wharton School

Abstract

As digitization of various economic and human behaviors become more prevalent, we examine whether data analysis capabilities can help with process- and innovation-oriented firm practices. Using detailed, firm level data on employee data analysis capabilities combined with a survey of organizational practices for 330 large firms, we find that while neither data skills nor process-related practices affect productivity directly, they have a substantial positive interaction. Specifically, firms with process-related practices receive a greater marginal benefit for the presence of or acquisition of data-related skills in their workforce. However, we do not find the same complementarities between data-related skills and innovation-oriented practices and at times the interaction can even be negative. These results are also unique to data-related skills and not IT skills generally. Overall these results highlight the potential tradeoffs of using data analytics at firm, similar to the tradeoffs between exploitation and exploration. We show that value of analytics (at least over our 1987-2007 sample period) is in enabling the use of internal and external data to drive process improvement, but data analytics capability does not immediately translate to greater innovation.

Introduction

As the availability of data about consumers, suppliers, competitors and partners proliferates, firms are beginning to realize the power of using data to make decisions. Data analytics has transformed many industries from manufacturing, to sports, to political consulting–instead of relying on the "gut-instincts" of middle managers, baseball scouts, or political pundits, decisions can now be based on and backed by large scale data analysis (Lewis 2004, Silver 2012). Part of this trend involves the exploitation of operational data that is provided as a by-product of the deployment of enterprise information systems (Aral et al. 2012, McAfee 2002). Social network information and mobile devices provide digital traces of individuals' activities and relationships that would be otherwise difficult or impossible to observe, and can be used to drive strategy when combined with complementary decision making and business practices (see a recent case example in Mims (2015)). Clickstream data and keyword searches collected from various digital media can offer information about consumer behavior in real-time, eliminating the need to conduct one-off surveys. These data can become even more powerful when combined, providing an unprecedented opportunity to tap into the pulse of economic activities as they are happening (Eagle and Pentland 2006, Wu and Brynjolfsson 2014).

Initial research into the value of data-driven decision making suggests that analytics can have considerable effects on firm performance (Brynjolfsson et al. 2011). However, as with many prior generations of information-technology related innovations, there is likely to be considerable variation in the benefits of these technologies (Hitt and Brynjolfsson 1996). One possibility is that different firm capabilities and organizational structures–such as skilled employees and decentralization of decision making–may affect the benefits of technology investments (Bresnahan et al. 2002, Tambe et al. 2012). Recent work by McElheran and Brynjolfsson (2015) shows that data-driven-decision making complements employee education and changes in the allocation of decision rights in manufacturing firms suggesting that this perspective is still relevant in an era of "big data." Alternatively, the variation in benefits may be due to how analytics complements different types of activities. From the early days of the IT and organizations literature, a common theme was that different information technologies can affect different types of decisions (see e.g., Leavitt and Whisler (1958), Zuboff (1988)). This theme also recurs in the more modern literature on IT-skills complementarities (Brynjolfsson and McAfee 2014, Levy and Murnane 1996). Thus, it is likely that the ability of firms to benefit from data analytics depends crucially on the capabilities that they have for conducting analytics, as well as benefit that analytics may have for important firm-level decisions.

In this paper, we take the perspective that there is large and principally exogenous shift in the amount of available data. Firms with the capability to exploit these data and decision-making tasks that can be leveraged by these data will benefit, while other firms may not. We operationalize the capabilities of a firm in data analytics as the proportion of employees throughout the firm who possess data-related skills. We then compare the benefits of these skills in firms that principally engaged in process-related decisions (the ongoing improvement of operational processes) versus innovation-related decisions (the creation and deployment of new products and services). We choose these two types of organizational practices because they are related to the exploitation and exploration in the strategy literature that has demonstrated the difficulty of doing both at the same time in a firm (Benner and Tushman 2003, March 1991). In particular, we hypothesize that data skills are more valuable in firms that are engaged in process-related decisions since these types of decisions are more amenable to analytical support. Because the level of data skills and the relative emphasis on process versus product innovation in firms likely varies for any number of reasons including idiosyncratic firm history, industry evolution, and product position, we will observe different combinations of data skills and decisions across firms in our sample which can be useful in identifying the performance effects.

Using data on firm-level characteristics from a large sample of online resumes combined with a firm-level survey of firm practices in 2008 for about 300 firms, we are able to determine the level of data skills and the absolute and relative emphasis of the firm on process versus innovation-related decisions. Incorporating these measures into standard models and data analysis methods from the IT value literature, we find that data skills appear to complement process-related decisions but not innovation-related decisions. Furthermore, we find some evidence that greater analytical capability and active use of analytics for innovation-related decisions may reduce performance. These results suggest data analytics skills are more likely to help exploiting incremental changes with existing business processes as opposed to facilitating innovation. To explore potential endogenous factors that could bias our results, we use instrumental variables computed from the data skills present in a firm's neighbors in its hiring network. Presumably, the data talent in network neighbors can lower the cost for the focal firm to acquire data talent but it should not directly affect the focal firm's productivity. Using the instrumental variable approach, we continue to find support for complementarities between data talent and process-related practices but not with innovation-related practices. Overall, our results provide some insight as to where analytics may have the most value, potentially explain why some firms may receive disproportionate benefits from external trends in the availability of consumer and commercial data, and provide support for recent work that suggests that certain types of tasks are likely to be increasingly driven by data and less by pure managerial judgment.

Theory and Literature

Data-driven decision-making has become more prominent in the recent years as data availability has grown. Brynjolfsson et al. (2011) is one of the first large-scale studies on the impact of data-driven decision-making. They show that firms adopt data-driven-decision making are more productive on average than firms that relies on experience and expertise. Saunders and Tambe (2013) also examined the productivity and market value impact for firms that embrace data use. They find data-intensive firms – as measured by discussion of data related topics in their annual reports—have higher productivity and market valuations. Recent work by McElheran and Brynjolfsson (2015) delves deeper into the impact of data-driven decision-making and find that manufacturing firms benefit from using data to make decisions when they adopt IT and modern organizational practices such as team-based decision making.

Our paper differs from these papers in two ways. First, we argue that in order to effectively make use of newly available external data, firms must have existing data talent who can make sense of the data and use insights learned from the data to make decisions. As data talent is often embodied within the company's workforce, we hypothesize that managerial intention to use data for decision-making is not enough; the firm also needs the capabilities to leverage data. One measure of how well a firm can utilize data is the number of employees with data analysis skills. Second, we argue while data can be extremely useful for decision-making, they may be more useful for certain firm decisions than for others. To the extent that the digitization of economic activity makes data more available for decision making, it is important to know both the advantages and the limitations of using data and the type of organizational practices that complement the use of external data.

Process-Oriented and Innovation-Oriented Firm practices

Since the mid-1990s, the IT literature has recognized that IT is a significant driver of productivity only when firms adopted a series of complementary organizational practices (Bresnahan et al. 2002, Hitt and Brynjolfsson 1996, Melville et al. 2004). The primary explanation for these complementarities is that they affect the ability of the firm to gather and act on information for operational decisions. For instance, decentralized decision making structures provide a greater ability to act on local, tacit knowledge by collocating decision rights with the necessary information. More recent studies have suggested the ability to gather and act upon external information has become an important complement to IT investment (Tambe et al. 2012). The emerging literature on corporate social media have shown that the ability to leverage the massive amount of unstructured data from social media can lead to a significant improvement in firm productivity (Hitt et al. 2015) especially when the ability to process information is dispersed throughout the firm.

Not all decisions are necessarily amenable to data-driven decision making. From the early days of the IT value literature it was recognized that different types of information (and the systems that provide and process this information) can be substitutes or complements to managerial judgment (Leavitt and Whisler 1958). Studies of the automation of work have found a similar theme (Zuboff 1988) – some decisions are easily automated while others are not. This research is closely related to work on skill-biased technical change, where different occupations can be differentially affected by the introduction of information systems and related organizational practices (Levy and Murnane 1996, Murnane et al. 1999). In some cases, routinized decisions can be replace entirely by information systems, while in other cases data can complement decision-making. These trends have received renewed interest as we are increasingly observing the effects of the "digitization of work" across increasingly wide domains (Brynjolfsson and McAfee 2014). While data based decision-making has nonetheless expanded into new domains, there are still areas that are not amenable to automation such as tasks involving creativity or insight. Regardless, at any given time and state of technology, certain types of decisions will be more amenable to data support than others.

To understand the advantages and disadvantages of using data for different types of corporate decisions we focus on two types of decisions: process-related decision and innovation-related decisions. We choose to compare process and innovation related practices because they mirror the dichotomy of exploration vs. exploitation strategies in firms. Prior literature has shown that it is often difficult for firms to both exploit and explore at the same time (Benner and Tushman 2003, March 1991). While data analysis capabilities can help a firm manage the large amount of data enabled through the recent digitization of various human behaviors, it is unclear whether these data capabilities can equally help firms exploit the incremental improvements in various existing business processes and at the same time explore innovation opportunities. Through this process, we explore the benefits and potential limitations of using data analysis skills as firms apply insights learned from digitization to various firm strategies.

Process-related decisions focus on the improvement of operational processes through the use of both internal and external information. Process improvement has a long history of being closely tied to IT usage such as the reengineering wave that follow the deployment of enterprise-wide IT (see e.g., Hammer (1990)). Using data to manage and refine processes has emerged as an important competitive advantage because it can help firms attain flexibility, speed, cost reductions and reduced risk (Barua et al. 2012, Mithas et al. 2011). Moreover, reducing the variance in production processes is critical for managing operations and the available data could play an important role in managing risks and improving efficiency (Bray and Mendelson 2012, Frei et al. 1999). The types of processes amenable to data-driven decisions have expanded. Similarly, hiring processes has been changed dramatically as firms embrace using data and HR analytics for both matching internal employees and external candidates to the right job (Bock 2015). Data availability and accuracy can be critical in ensuring forecasting and streamlining supply chain management processes for retailers (Fisher et al. 2000). Data management and information capabilities can also improve business processes for outsourcing companies (Mani et al. 2010). Traditional industries such as agriculture have also been transformed because of data. Geo-location capability and satellite data services can improve the precision of planting and chemical application and real-time data can improve forecasts of yield and quality which improves profitability (Mucherino et al. 2009).

As data availability grows from digitization of various human and economic behaviors, firms across various industries can benefit substantially from using these data. Often, they capture traces left from various decision-making and economic behaviors as the direct result of existing business processes. Analyzing these digital traces and ultimately translating insights learned from data into strategies are valuable for improving the efficiency of existing business processes. The data analytics capabilities are especially valuable given the bounded rationality of individual decision-makers that cannot possibly retain the vast amount of dynamic information generated every day. Having data analysis skills that could digest the data into actionable insights can thus help stakeholders make critical business decisions.

H1: Data skills and process-oriented firm practices are complements for improving firm productivity.

Analytics can also potentially support the innovative process, but here the relationship is less clear. On the one hand, there is a substantial increase in the amount of available data about customer preferences and behaviors through various sources (e.g., search data, social media, product reviews etc.). Because these data are more detailed than ever before, firms are able to dissect consumer preferences and needs at a granular level, allowing them to develop new products to cater to increasingly narrow market segments. Some research has also tied IT investment to increased trademark activity which is at least partially indicative of product innovation through new product introduction (Gao and Hitt 2012). However, capturing consumer even at the finest level often leads to evolutionary rather than revolutionary innovations (Moon 2010). The innovative process relies much more on tacit knowledge and other types of information that is less amenable to quantitative analysis (Henderson and Clark 1990, Nonaka and Von Krogh 2009). Avery and Norton (2014) noted what people revealed about themselves on paper or online could be quite different from their true motivations. As the result, consumer research often reveals the rationales behind the purchase that are formed ex-post to justify the decisions but may not capture the true motivations behind the actual purchase decision. Anecdotal evidence also suggests that some leading innovators, such as Steve Jobs, deliberately limit the use of analytics in the creative process. Notably, Jobs refused to conduct any marketing research when creating the first iPad. Given the large amount of negative sentiment and prior history of failed tablet computing innovations, it is quite possible that the iPad would have never been created (in contrast to the negative predictions, 1 million iPads were sold in the first month of availability and more than 200 million to date).[1] Similarly, because market research is very good at determining consumer preferences among existing products, it is not necessarily useful for evaluating future behavior for products that do not yet exist. This is precisely because consumers often articulate their preferences based on experiences with existing products in the marketplace, but they are limited in their imagination of what the product could be. As the result, learning from data about consumer behaviors and preferences often leads to imitations and incremental improvements on existing products as opposed to creating revolutionary innovations. This exploitation for incremental innovations often leads to feature wars among competitors in search of local maximum as opposed to finding exploratory and radical innovation (Benner and Tushman 2002, Lieberman and Asaba 2006).

Furthermore, mining for common patterns in consumer data, regardless how detailed observations they can be, often serve the purpose of confirming manager's pre-existing beliefs about consumers rather than explore new hypothesis (Martin 2009). Thus, the exogenous shift in data availability as well as the possessing the data analytical talent would not benefit innovation when they are used to confirm existing biases and justify decisions ex-post. While experimental methods may help close this gap – for instance, Capital One routinely conducts experiments to determine new credit card offerings (Clemons et al. 1995) but these experiments are very costly for most firms, except those in Internet retail or online services where A-B testing and experimentation has become commonplace (Clay et al. 2001, Einav et al. 2011). Martin (2009) suggests actively looking for anomalous data that challenges accepted explanations require richer data about a limited number of exceptional customers, rather from the mass data about consumer behaviors generally as is typical for online analytics (Avery and Norton 2014). Thus, while it is possible that analytics can support innovation related decision, it is likely to be considerably less useful for firms engaged in product innovation without significant organizational redesign.

H2: Data skills are not necessarily complementary to innovation-oriented practices.

Measuring the Impact of Data Analysis Skills

Productivity

To understand how data analysis skills can affect work performance, we leverage techniques developed in the IT value literature to determine the marginal impact of information technology investment after accounting for various firm inputs and external factors such as time and industry. Instead of measuring the performance effect of IT investments, we focus on measuring the effect of having employees with data analysis skills, as well as the type of organizational decisions these skills are particularly suited to enable.

While there are a number of approaches to measure the marginal effect of information technology on firm performance, we choose to use the multi-factor productivity framework, specifically the Cobb-Douglas function that is the most commonly used framework in estimating IT value. Using the this framework, we relate firm output such as sales or value-added, to various firm inputs such as labor, capital and materials after controlling for industry and temporal variations. The residual of this equation can be interpreted as

[1] http://www.statista.com/statistics/269915/global-apple-ipad-sales-since-q3-2010/

the multi-factor productivity and we are interested in exploring if data analysis skills is related to this measure of productivity as shown in the following equation.

$$\ln(sales)_{it} = \beta_0 + \beta_1 \ln(m)_{it} + \beta_2 \ln(k)_{it} + \beta_3 \ln(l)_{it} + \beta_4 \mathrm{Data}_{it} + controls + \epsilon$$

Thus, we relate sales to production inputs such as materials expense (m), physical capital stock (k) and labor (l, measured as the number of employees) as well as controls for industry(i) and time (t). Here and throughout we will use italics to designate specific regression measures (see Tables 1 and 2 for further detail). To this standard framework we add a measure of data analysis capability (Data), which is the number of employees that possess data-related skills. Since we are focusing on labor-related issues, we also control for average education, age, and gender of the employees to prevent our results from being confounded with labor quality. Education provides a measure of initial general human capital stock, while the combination of age and gender provide a sense of on-the-job human capital accumulation through experience (recognizing the relationship can be different for men and women). Our panel analysis includes the performance and labor metrics on a 20-year panel from 1987 to 2007.

While prior work has shown that data use and data-driven decision making can positively contribute to productivity (Brynjolfsson et al. 2011, McElheran 2011), we focus on the specific type of organizational decisions that can best benefit from data. In particular, we examine firms engaging in more process-oriented decisions when they make a greater emphasis on process management related practices (Process), or innovation-related decisions when innovation-oriented practices are more important or prevalent (Inno). These practices are measured separately so it is possible for some firms to be engaged in both or neither (see Table 1). To examine complementarity with data skills, we augment the production function with measures of these constructs as well as their interaction with data skills:

$$\begin{aligned}\ln(sales)_{it} = {} & \beta_0 + \beta_1 \ln(m)_{it} + \beta_2 \ln(k)_{it} + \beta_3 \ln(l)_{it} + \beta_4 \mathrm{Data}_{it} + \beta_5 \mathrm{Process}_i + \beta_6 Inno_i + \beta_7 \mathrm{Data}_{it} \mathrm{X}\ \mathrm{Process}_i \\ & + \beta_8 \mathrm{Data}_{it}\ \mathrm{X}\ \mathrm{Inno}_i + co\square trols + \epsilon\end{aligned}$$

Note that data skill varies both by time and across firms, while we have a single cross-section of organizational practices.

Data and Measurement

Business Practices

The business practices variables are constructed using a survey administered to senior human resource managers and chief information officers from large public traded firms in 2008. The survey is a joint effort between a two academic institutions and McKinsey and Company. In total, 331 firms responded to the survey. The survey mainly asks about various business practices as well as usage of various information systems. The questions were drawn in part from two previous surveys on IT usage and workplace organization practices conducted in 1995 and 2001 (Brynjolfsson and Hitt 1996, Brynjolfsson et al. 2011), but added questions about innovative and process-oriented business practices, as well as general decision-making practices. To explore what type of organizational practices that data analysis skills can complement, we use survey responses to construct two organizational practices scales. We measure process-oriented firm practices as a composite scale of eight questions (scored on a 1-5 scale) about the importance of business processes to firm value and how these processes are implemented. Four of these questions focus on the extent of which refining and improving existing processes is important to the firm (P1-P4). The next four questions (P5-P8) ask about the specific process-related practices that the firm uses, such as monitoring and analyzing processes. The scale is constructed as the standardized sum of the standardized values of the scale components.

$$Process = Norm(Norm(p1) + Norm(P2) + \cdots + Norm(P8))$$

Correlations among the individual constructs within this scale are all positive and the Cronbach's alpha is 0.712. To test the validity of including all eight measures into a single component, we separately introduced the measures into our main regression and find we cannot reject the hypothesis that all eight practices have the same coefficient.

We construct a second scale measuring innovation-oriented practices using responses to seven survey questions. All these questions ask about the general innovative practices at the firm, such as how long it takes to introduce a new product and develop new capabilities, as well as whether the firm is a leading adopter of new technologies. The first two questions (I1-I2) measure how fast the focal firm can introduce an innovative product or service, and the next three questions (I3-I6) measure how important innovation is for the firm. The last two questions (I6-I7) measure the likelihood of a firm to be the leading edge adopters of technology. The same approach was used to construct the scale as the process-oriented measure.

The Cronbach's alpha of this scale is 0.475. The relatively low value reflects the multidimensionality of innovative practices—firms adopting any one practice do not necessarily adopt all of the others. Firm and industry characteristics can also lead to divergent practices. For example, for a leading software engineering firm may find being a lead adopter of technologies to be important since the reward for being a lead adopter is high for the high tech industry. On the other hand, manufacturing firms may find innovative production methods and business capabilities to be more important. Our goal in this paper is not to identify which practices are the most beneficial but to evaluate the overall extent to which a firm engages in innovative practices. To test the validity of having all seven survey responses into a single construct, we also separately introduce each measure into our main regression and find that we cannot reject the null hypothesis that all seven practices have the same regression estimates. Consequently, we choose to combine all seven measures into a single innovative practice measure. The survey questions and the summary statistics of the individual questions and the scales as a whole are provided in Table 1.

Data Analysis Skills

In order to test for how data skills among employees affect firm productivity as well as how they influence firm's innovation and process-oriented practices, we use a database consisting of more than 6 million individual resumes in 2007. These data are similar to other large sample resume datasets used for prior work in IT value and technology diffusion [see e.g., Tambe and Hitt (2013) for a more detailed discussion of the advantages and limitations of these datasets generally]. Using natural language processing techniques, we identify the skills of each employee as reflected in the free text part of their resume as well as the words that appear in their job titles. The text mining algorithm maps free text responses to a taxonomy of skills (this includes direct matches on keywords like "analytics" as well as inferences that can be made by phrases such as "regression modeling"). We then classify the skills within the taxonomy into data-related and non-data related skills. Overall, the distribution of data analysis skills across different areas of the firm is broad and spans multiple business areas. They include consultant, financial analyst, systems engineer, customer service specialist, program manager, and systems analyst for instance. For each employee in our data, we identify whether a skill is present in his/her resume and aggregate the total number of employees with the skill for each firm. Because we have the career history of each person in the resume, we can approximate the timing of when the person acquired and actively use the data analysis skills we identify.

IT Skill

As a control and a falsification test, we also construct a measure of IT skills using a similar approach to Tambe and Hitt (2013). This method focuses on IT-related job titles (and does not rely on the skills taxonomy described earlier). IT workers are identified either by titles that are clearly related to information technology (for example, software engineer, systems analyst, programmer analyst) or contain keywords that suggest the employees is an IT worker (for example, computer, website, software). We conducted various alternative measures of identifying IT or job skills such as using the content of each job description, but they do not qualitatively change our analysis. Similar to data skills, we constructed the level of IT skills each firm has over a 20-year panel and match it to the firm's performance measures. This measurement of IT skills has been verified to accurately capture the raw IT talent in each firm (Tambe and Hitt 2013, Wu et al. 2014). In addition, these measures perform similarly, if not better, than more traditional capital-based measures of IT investment (Tambe and Hitt 2013) in productivity measurement frameworks.

Production Inputs and Performance

We use Compustat Industrial Annual files from 1987 to 2007 to measure other production inputs such as physical assets, employees, and sales. Materials were calculated sales minus the sum of operating income and labor expense. When labor expense is not available, we estimate it by multiplying the total number of employees and the industry average wage for the most disaggregated industry data available. These techniques are consistent with approaches used in the R&D and IT-productivity literature. The summary statistics for the financial variables and data skills are shown in Table 2.

Results

The descriptive statistics for our main variables of interests are shown in Table 1 and Table 2. Most of the business practice questions were captured on a 5-point or a 7-point Likert scales with a mean on the order of 3-4 and a standard deviation of approximately 1. Figure 1 and 2 shows the distribution of the process and innovation related practices. The distribution for process-oriented practices is symmetric around zero whereas the distribution for innovation-oriented firm practices is skewed toward the right; the mode in the histogram for innovation practices is greater than the mean. Table 2 shows the summary statistics for the performance metrics of the firms and Table 3 shows the pairwise correlation table among the variables of interest. While innovation and process-oriented practices are correlated, the magnitude of the coefficient is relatively low at 0.363. This shows that some firms clearly put process improvement above innovation while other firms put innovation at a higher priority. Figure 3 shows the distribution of firms along the process and innovation axes. Firms are fairly evenly distributed in the four quadrants.

The correlation between having IT skills and having data skills is also positive but the magnitude is also low at 0.121, indicating the two skills are distinct. Firms that were historically high in IT may or may not recognize the importance of having data skills or have made the necessary investment to use data for decision-making. This will tend to lower the estimates of correlation but it will strengthen the power of the performance regressions.

The primary results regarding the relationship among data analysis skills, innovation or process oriented firm practices and productivity are shown in Table 4. In Column 1 of Table 4, we estimate the effect of having data analytics talent in the firm using a fixed-effects model. Interesting, data skills by itself does not contribute to firm productivity. While the estimate is positive, it is not different from zero. In Column 2 we examine whether firms using more process-oriented practices are more productive. We choose to use an OLS model with (firm-level) clustered standard errors because the fixed-effect model could not separately estimate the effect of process-related practices because our firm practices are collected using a single survey conducted in 2008. In the OLS estimate, we show that process-oriented firms are not necessarily more productive as the coefficient is not statistically different from zero. In Column 3, we estimate whether having data talent embodied in a firm's workforce can help firms that are also process-oriented. The OLS estimate with clustered standard error shows while process-oriented practices and data skills alone do not directly contribute to firm productivity, their interaction is positive. On average, a one-standard deviation increase in data skills helps process-oriented firm improve productivity by 3.34%. This is after controlling for industry, year, firm age, and the average education of employees in each firm. The result also continues to hold after controlling for IT labor. This shows that the variation in productivity due to the data-process complementarity is not confounded with ordinary returns to IT investments.

While we prefer the explanation that the complementarities between data skills and process-oriented practices generate superior firm performance, there are at least two endogenous factors that could lead to a positive bias on our estimate. First, high performing firms can have slack resources that allow them to invest heavily in data talent acquisition, which could lead to a reverse causality between performance and data talent for process-oriented firms. Second, there could be omitted variables such as higher management quality also attracts higher quality labor that lead to an upward bias. To address these problems, we treat data talent as endogenous and use three instruments to correct the potential biases: 1) the average percentage of employee with data skills in a firm's existing hiring network; 2) the average number of employees with data skills in a firm's existing hiring network; 3) the sum of all data employees in a firm's immediate hiring network. All three instruments are derived from a firm's immediate hiring network, which can be used as a proxy to measure the ease of access to existing data talent. The hiring network is constructed by examining employee flows across firms using similar techniques to those in (Tambe and Hitt 2013). The first stage concentration parameter is over 100, indicating that weak instruments are not an issue. In Column 4, we show the 2SLS result using the three instrumental variables and their associated interaction term. Again, we observe that process-oriented firm practices positively interact with data skills and the estimate is similar to the OLS estimate (β=.0446, ρ<0.1). We also use the SYS-GMM approach (Arellano and Bond 1991) and the associated instruments (lagged differences and levels) to control for the potential endogenous factors arise from production inputs such as materials, capital and labor. Using two-period lags as well as the network-based instruments for data talent, we estimate a two-step GMM model with robust standard errors in Column 5. We find similar results as in our OLS and 2SLS estimate; there is a positive interaction between data skills and using process-oriented practices (β=.0757, ρ<0.01).

Another potential source of endogeneity is that firm practices such as process-oriented and innovation-oriented practices may be endogenous. Because our survey data is cross-sectional, we cannot directly assess the level of practices before and after a change in firms' talent acquisition in data skills. However, we take advantage of the fact that organizational practices are often quasi-fixed (Bresnahan 2001, Brynjolfsson and Hitt 1996, Levy and Murnane 1996, Milgrom and Roberts 1990). Thus, our regression results can be interpreted as assessing whether pre-existing firm differences in process and innovation-oriented practices influence the productivity return from acquiring data talent. Under the quasi-fixed assumption, firms that already have implemented process-oriented firm practices are more likely to acquire data talent because it can enhances the effectives of these organizational practices and these firms receive greater benefit than other firms for these investments. Overall, these results show that when a firm is process-oriented, having data talent can particularly help the firm to become more productive. As the consumer behaviors, choices and other traces left behind by various processes are digitized, firms can derive substantial benefits from using these data to improve their existing business processes. Having existing data skills can thus help firms take advantage of the data and derive insights to improve existing processes. Overall, these results support hypothesis H1.

While the availability of a large amount of data about various aspects of firms' operations readily supports process-oriented firms, it is unclear whether innovation-oriented firms can also benefit from types of data that are amenable to digitization given the current state of the technology. To test this hypothesis, we perform a parallel analysis using our measure of innovative practices. In Column 6, we show that adopting innovative-oriented practices alone does not seem to provide any edge in the return to firm productivity. Furthermore, in contrast to process-oriented firms that benefits from having data skills, innovation-oriented firms do not seem to benefit from having data skills. In fact, having data skills seem to worsen the productivity for firms that use innovation-oriented practices. As shown in Column 7, the interaction between data skills and innovation practices is negative (β=-.0567, ρ<0.1). In Column 8, the 2SLS model that uses data skills of neighboring hiring firms as an instrument for the firms' own data skill show the same pattern. Again we see that the interaction between data skills and innovation-oriented practices is negative. GMM estimates using both external instrumental variables and lags of the independent variables show a very similar estimate to the 2SLS model in Column 9 (β=-.0815, ρ<0.1). The OLS, 2SLS and GMM model consistently show that innovation-oriented practices and data skills negatively interact to affect firm performance at a borderline level of significance. It appears having data analytics skills to take advantage of the newly available data can potentially reduce innovation.

Finally, we include process-oriented practices and innovation-oriented practices as well as their interactions with data skills in a single model. In the OLS estimate with clustered standard errors (Column 10), the interaction between process-oriented practices and data skills are positive and the interaction between data skills and innovation-oriented practices is negative, but they are not statistically significant. However, the interactions become statistically significant using the 2SLS model and the GMM model with external instruments and internal lags of the independent variables (Column 11 and 12). On average, a one standard deviation increase in data skills is associated with 3.80 to 4.76% improvement in productivity when firms are using process-oriented organizational practices. However, interaction between data skills and innovation-oriented firm practices failed to be statistically significant, albeit the coefficients are still negative. This may be due to our limited sample, measurement error or other types of unobserved heterogeneity associated with innovative practice.

Data skills vs. IT skills

To ensure that data skills are not simply capturing the effects of IT labor, we reproduce our results in Table 4, replacing data skills with IT skills. Instead of finding any complementarities with process-oriented firms, we show that IT skills are not complementary to process-oriented firm practices, nor are they complementary to innovation-oriented practices (Table 5). As shown in Columns 1-6, none of the interactions with IT skills are statistically significant. These results suggest that data skills are distinct from IT skills at least as we measure them. While earlier work has shown that IT skills are complementary to organizational practices such as decentralization and team work, data skills seem to be more complementary to process-driven firm practices.

Conclusion and Discussion

By using existing data talent as a proxy for data analysis skills at a firm, we show that data analytic skills in a firms workforce by itself does not seem to contribute to firm productivity, at least in the period from 1987 to 2007. However, we do find a significant complementarity between process-oriented firm practices and data skill suggesting that newly available data supports oriented decision-making. We do not find similar results for innovative activity, and our results appear to be related specifically to data analysis capabilities not general IT investment.

It may be that the types of data available during our study period (through 2007) are especially beneficial to firms engaged in process improvement. Here, having a workforce with strong analytical skill enables the internal and external data collected by enterprise systems (e.g. ERP, CRM, SCM) to be used for process optimization. We find no evidence, however, that general IT investments are particularly suited to one type of decision or another perhaps because IT investment for operational and innovative support is mature and thus does not contribute incremental to firm performance variation. This would be consistent with many of the studies that showed substantial benefits of ERP adoption and in the late 1990s when ERP systems were still diffusing across large firms; today, most firm whose structure and activity is amenable to ERP have likely already adopted these systems.

We also do not find either IT or data skills to complement the innovative process. In fact, our results show that data skills and the propensity to use data could potentially hurt firms with innovation-oriented practices though we did not show this consistently across our models. This may be due to the fact that innovative activity is not currently amenable to data analysis or that the types of data available throughout most of our sample period (which predates the large development of social media) do not provide the necessary information. Furthermore, using consumer behavioral data without observing the consumers' intentions and motivations could often lead to incremental innovations, because consumers have limited ability to express preferences over products or product attributes that do not yet exist. However, we also observed a substantial heterogeneity in the return to using data and innovation-oriented practices. Perhaps, some unobserved capabilities allow firms to better use the data for appropriate type of innovations.

Overall, these results show that the massive increase in the availability of consumer data through the digitization of consumer behaviors is more closely aligned with process improvements rather than product innovation. This may suggest that the scarce resources focused on "big data" analytics are best directed at operational improvements rather than new product development. In addition it suggests that the type of analytics that were possible in the pre-social media era were not sufficient rich to enable innovative activity. Recent results suggest that social media and data analytics may be complementary (Hitt et al. 2015) perhaps indicating that richer data and new techniques common to this era (e.g. marketplace experiments, real-time response measurement) may begin to support the innovative process as well.

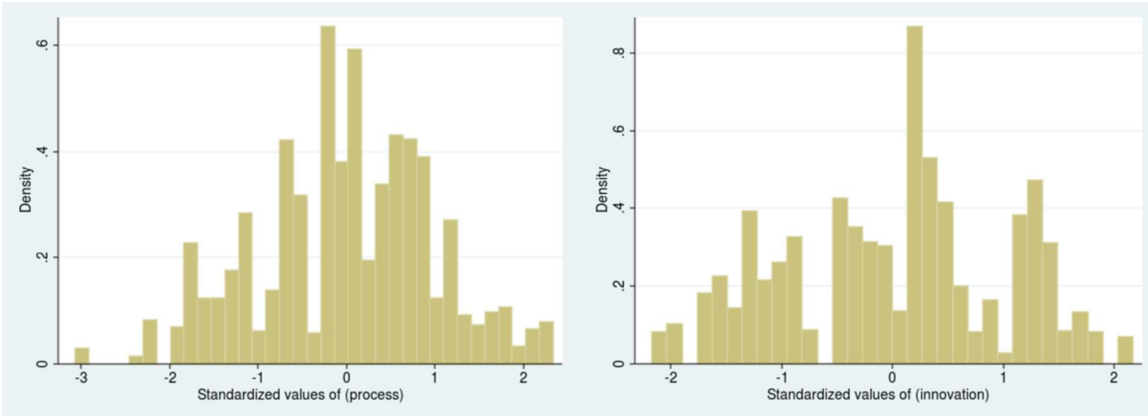

Figure 1: Distribution of Process and Innovation-oriented Organizational Practices

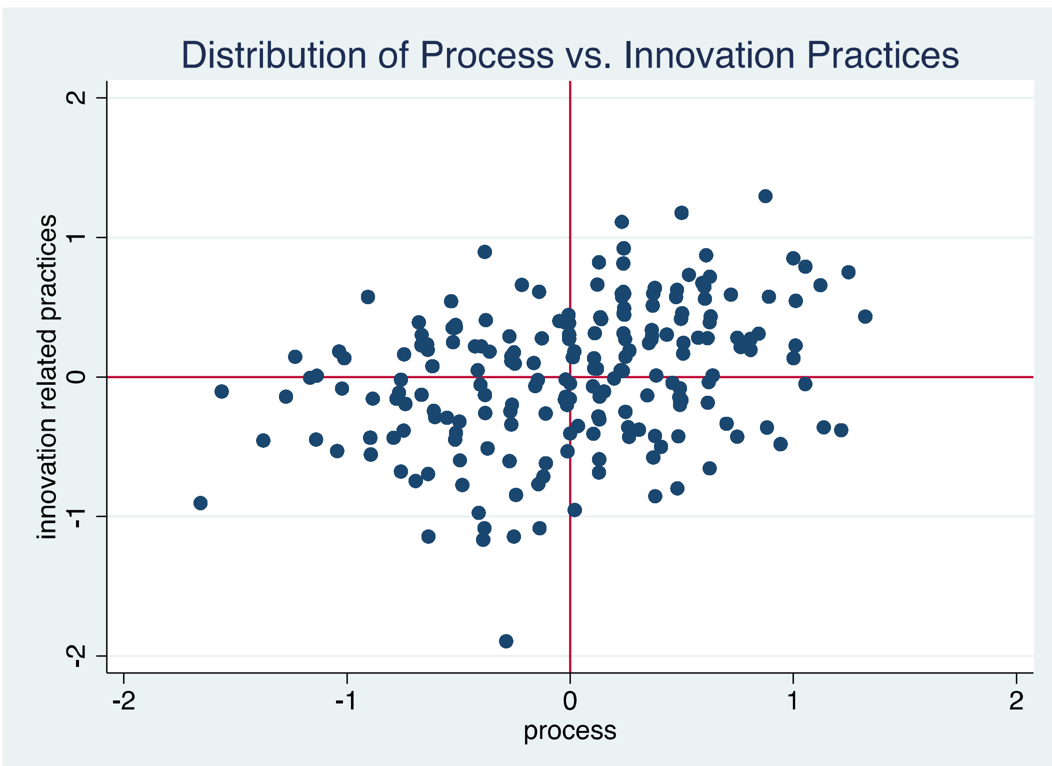

Figure 2: Distribution of Firms on the Process and Innovation axis

Table 1: Process-oriented and Innovation-oriented Variables						
Process-oriented Org Practices						
	Survey Question	Obs.	Avg.	Std. Dev.	Min	Max
P1	To what extent do the following statement describe the work practices and environment of your entire company: We have a strong ability to **make incremental changes or improvements to our business processes**.	3593	3.58 2	.880	1	5
P2	To what extent do the following statement describe the work practices and environment of your entire company: Our company **stresses operational excellence** (efficient execution) over innovation.	3586	3.28 6	1.069	1	5
P3	Please list the most important core activities for your primary business: **Process development, process quality, process management and improvement**	3416	.052	.222	0	1
P4	Currently, how effective is your IT organization in proactively engaging with business leaders to **refine existing processes and systems**.	3552	3.04 5	.988	1	5
P5	To what extent do the following statement describe the work practices and environment of your entire company: We have a cross-functional process orientation.	3570	3.15 4	1.025	1	5
P6	Analysis and improvement are part of everyone's job	3557	3.55 1	1.021	1	5
P7	We rigorously track and ensure capture of the benefits or returns specified in business cases	3558	2.85 7	.967	1	5
P8	We monitor performance of activities through a set of well-defined metrics	3577	3.35 7	.992	1	5
	Process = Norm(Norm(P1) +...+Norm(P8))	3628	0	1	-2.890	2.292
Innovation-oriented Org Practices						
	Survey Question	Obs.	Avg.	Std. Dev.	Min	Max
I1	On average, how long does it take your organization to perform the following activities: **Design, create and introduce a new product or service** after approval.	3244	4.563	1.400	1	7
I2	On average, how long does it take your organization to perform the following activities? **Introduce a new production technology or method.**	3066	4.552	1.172	2	7
I3	List the most important core activity for your primary business unit: **product development, product design, innovation**	3416	.167	.373	0	1
I4	To what extent do the following statement describe the work practices and environment of your entire company: Our company **stresses innovation** (efficient execution) over operational excellence.	3586	2.714	1.069	1	5
I5	To what extent do the following statements describe the work practices and environment of your *entire company:* We encourage experimentation to continually improve our offerings	3540	3.048	1.029	1	5
I6	To what extent do the following statements describe the work practices and environment of your *entire company:* We are usually the **leading-edge adopter** of new technologies in our industry.	3545	2.623	1.178	1	5
I7	How would you generally characterize your company's approach to use technology in support of business capabilities? (Choose one) (a) Leading edge adopter; (b) Fast follower; (c) Mature adopter; (d) Late adopter	4273	2.489	.830	1	4
	Innovation = Norm(Norm(I1) +...+Norm(I9))	3841	0	1	-4.481	3.120

Table 2. Descriptive Statistics					
Variable	Obs.	Avg.	Std. Dev.	Min	Max
Sales (MM$)	5609	3592.902	13552.39	0	458361
Materials (MM$)	5057	237.061	11237.19	.0364	389650
Capital (MM$)	5252	2951.685	11328.45	0	235278
Employee (M)	5335	13.894	27.689	0	553.6
Capital intensity	4915	1.779	2.328	0	56.644
Tobin's Q	3139	1.831	1.402	.347	34.534
Data Skills	4248	21.343	70.609	0	1416
IT Skills	4081	1463.008	4604.571	1	75501

Table 3. Correlations									
	Sales	Material	Capital	Employee	Data Skill	IT Skills	Process	Innovation	DDD
Sales	1								
Material	0.9958	1							
Capital	0.9461	0.9351	1						
Employee	0.555	0.5112	0.5353	1					
Data Skills	0.1198	0.0582	0.1122	0.3778	1				
IT Skills	0.2403	0.185	0.2086	0.5437	0.4503	1			
Process	0.0695	0.056	0.0584	0.1781	0.1125	0.1254	1		
Innovation	-0.107	-0.0955	-0.1323	-0.0443	-0.0977	-0.1206	0.2964	1	
DDD	0.1112	0.103	0.1134	0.1101	0.1147	0.0996	0.5554	0.0946	1

Table 4. Relationship between data skills, firm practices and productivity

DP=lnSales	(1) FE	(2) OLS	(3) OLS	(4) 2SLS	(5) GMM/IV	(6) OLS	(7) OLS	(8) 2SLS	(9) GMM/IV	(10) OLS	(11) 2SLS	(12) GMM/IV
ln(materials)	0.674***	0.606***	0.607***	0.607***	0.670***	0.605***	0.607***	0.607***	0.701***	0.607***	0.607***	0.626***
	(0.0671)	(0.0383)	(0.0381)	(0.0376)	(0.0885)	(0.0388)	(0.0388)	(0.0383)	(0.0934)	(0.0386)	(0.0380)	(0.0463)
ln(capital)	0.101**	0.133***	0.130***	0.130***	0.0973	0.131***	0.127***	0.127***	0.0738	0.126***	0.125***	0.144***
	(0.0512)	(0.0237)	(0.0238)	(0.0236)	(0.0658)	(0.0239)	(0.0242)	(0.0241)	(0.0748)	(0.0241)	(0.0238)	(0.0392)
ln(employees)	0.284***	0.237***	0.239***	0.243***	0.258**	0.240***	0.240***	0.245***	0.249**	0.244***	0.248***	0.227***
	(0.0793)	(0.0262)	(0.0261)	(0.0256)	(0.117)	(0.0267)	(0.0264)	(0.0261)	(0.109)	(0.0265)	(0.0262)	(0.0429)
Processes-oriented Practices		-0.00364	-0.00741	-0.00777	-0.0237					-0.00542	-0.00547	-0.0134
		(0.0176)	(0.0182)	(0.0187)	(0.0207)					(0.0199)	(0.0198)	(0.0200)
Data skills	0.0294		-0.00912	-0.0342	-0.0765*		-0.00180	-0.0376	-0.0572	-0.0258	-0.0535	-0.0642
	(0.0421)		(0.0301)	(0.0388)	(0.0448)		(0.0250)	(0.0368)	(0.0444)	(0.0368)	(0.0444)	(0.0416)
Data X Process			0.0334*	0.0446*	0.0757***					0.0268	0.0380*	0.0476**
			(0.0178)	(0.0266)	(0.0240)					(0.0191)	(0.0198)	(0.0210)
Innovation-oriented practices						-0.0163	-0.0130	-0.0167	-0.0326	-0.0124	-0.0148	-0.0147
						(0.0253)	(0.0249)	(0.0250)	(0.0302)	(0.0277)	(0.0277)	(0.0286)
Data X Innovation							-0.0567*	-0.0816*	-0.0815*	-0.0449	-0.0560	-0.0413
							(0.0321)	(0.0447)	(0.0461)	(0.0325)	(0.0367)	(0.0337)
Constant	0.255	1.232***	1.242***	1.522***	1.029***	1.246***	1.261***	1.538***	1.083***	1.255***	1.544***	1.089***
	(0.674)	(0.0819)	(0.0864)	(0.0958)	(0.340)	(0.0863)	(0.0896)	(0.0976)	(0.316)	(0.0905)	(0.0993)	(0.122)
Observations	2,282	2,282	2,282	2,282	2,282	2,263	2,263	2,263	2,263	2,263	2,263	2,263
R-squared	0.981	0.968	0.968	0.968		0.967	0.967	0.967		0.967	0.967	
Industry	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: Instrumental variables included the GMM/IV estimation are 1) the average ratio of data employees in the neighboring firm; 2) average of the total number of data employees in the neighboring firm 3) the sum of all data employees in the neighboring firm. We also used one and two period lags for the GMM estimation as well. Robust standard errors in parentheses *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 5. Relationship between IT skills, firm practices and productivity

DP=lnSales	(1) FE	(2) OLS	(3) OLS	(4) GMM	(5) OLS	(6) OLS	(7) GMM/IV	(8) OLS	(9) GMM/IV
ln(materials)	0.671***	0.606***	0.607***	0.636***	0.605***	0.606***	0.628***	0.606***	0.602***
	(0.0675)	(0.0383)	(0.0378)	(0.0516)	(0.0388)	(0.0384)	(0.0447)	(0.0383)	(0.0601)
ln(capital)	0.103**	0.133***	0.130***	0.157***	0.131***	0.128***	0.161***	0.128***	0.182***
	(0.0517)	(0.0238)	(0.0236)	(0.0467)	(0.0239)	(0.0240)	(0.0408)	(0.0239)	(0.0360)
ln(employees)	0.277***	0.237***	0.231***	0.161***	0.240***	0.231***	0.164***	0.234***	0.184***
	(0.0805)	(0.0262)	(0.0262)	(0.0521)	(0.0267)	(0.0271)	(0.0412)	(0.0266)	(0.0449)
Processes-oriented Practices		-0.00354	-0.00661	-0.00557				-0.00569	-0.0115
		(0.0176)	(0.0180)	(0.0220)				(0.0198)	(0.0232)
IT labor	0.0537		0.0515	0.0156		0.0589	0.0371	0.0318	0.0443
	(0.0722)		(0.0408)	(0.0706)		(0.0556)	(0.0844)	(0.0493)	(0.0972)
IT skill X Process			0.0459	0.0857*				0.0423	0.0188
			(0.0354)	(0.0474)				(0.0382)	(0.0575)
Innovation-oriented practices					-0.0163	-0.0103	-0.00374	-0.00770	0.00103
					(0.0253)	(0.0254)	(0.0297)	(0.0280)	(0.0310)
Labor IT X Innovation						-0.0533	-0.143	-0.0449	-0.0932
						(0.0735)	(0.128)	(0.0796)	(0.113)
Constant	0.307	1.232***	1.269***	1.017***	1.246***	1.271***	0.899***	1.271***	0.956***
	(0.688)	(0.0819)	(0.0838)	(0.260)	(0.0864)	(0.0902)	(0.293)	(0.0889)	(0.345)
Observations	2,277	2,277	2,277	2,277	2,258	2,258	2,258	2,258	2,258
R-squared	0.981	0.968	0.968		0.967	0.967		0.967	
Industry	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

References

Aral S, Brynjolfsson E, Wu L (2012). Three-way complementarities: Performance pay, human resource analytics, and information technology. *Management Science*. 58(5) 913-931.

Arellano M, Bond S (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The review of economic studies*. 58(2) 277-297.

Avery J, Norton M (2014). Learning From Extreme Consumers.

Barua A, Mani D, Mukherjee R (2012). Measuring the business impacts of effective data. *Report accessed at* http://www. *sybase. com/files/White_Papers on Sep*. 15 2012.

Benner MJ, Tushman M (2002). Process Management and Technological Innovation: A Longitudinal Study of the Photography and Paint Industries. *Administrative Science Quarterly*. 47(4) 676-706.

Benner MJ, Tushman ML (2003). Exploitation, exploration, and process management: The productivity dilemma revisited. *Academy of management review*. 28(2) 238-256.

Bray RL, Mendelson H (2012). Information transmission and the bullwhip effect: An empirical investigation. *Management Science*. 58(5) 860-875.

Bresnahan TF (2001). The mechanisms of information technology's contribution to economic growth. *Institutions, Innovation and Growth: Selected Economic Papers* 135-137.

Bresnahan TF, Brynjolfsson E, Hitt LM (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The Quarterly Journal of Economics*. 117(1) 339-376.

Bresnahan TF, Brynjolfsson E, Hitt LM (2002). *Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence*.

Brynjolfsson E, Hitt L (1996). Paradox lost? Firm-level evidence on the returns to information systems spending. *Management science*. 42(4) 541-558.

Brynjolfsson E, Hitt L, Kim H (2011). Strength in Numbers: How does data-driven decision-making affect firm performance?

Brynjolfsson E, McAfee A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Clay K, Krishnan R, Smith M (2001). The great experiment: Pricing on the Internet. *The Heinz School, Carnegie Mellon*.

Clemons EK, Thatcher ME, Row MC (1995). Identifying sources of reengineering failures: A study of the behavioral factors contributing to reengineering risks. *Journal of Management Information Systems* 9-36.

Eagle N, Pentland A (2006). Reality mining: sensing complex social systems. *Personal and ubiquitous computing*. 10(4) 255-268.

Einav L, Kuchler T, Levin JD, Sundaresan N (2011). *Learning from seller experiments in online markets*. National Bureau of Economic Research.

Fisher ML, Raman A, McClelland AS (2000). Rocket science retailing is almost here-are you ready? *Harvard Business Review*. 78(4) 115-124.

Frei FX, Kalakota R, Leone AJ, Marx LM (1999). Process variation as a determinant of bank performance: evidence from the retail banking study. *Management Science*. 45(9) 1210-1220.

Gao G, Hitt LM (2012). Information technology and trademarks: Implications for product variety. *Management Science*. 58(6) 1211-1226.
Hammer M (1990). Reengineering work: don't automate, obliterate. *Harvard business review*. 68(4) 104-112.
Henderson RM, Clark KB (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative science quarterly* 9-30.
Hitt L, Jin F, Wu L (2015). *Data Skills and Value of Social Media: Evidence from Large-Sample Firm Value Analysis*. New Zealand.
Hitt LM, Brynjolfsson E (1996). Productivity, business profitability, and consumer surplus: three different measures of information technology value. *MIS quarterly* 121-142.
Leavitt HJ, Whisler TL. (1958). *Management in the 1980's*. November.
Levy F, Murnane RJ (1996). With what skills are computers a complement? *The American Economic Review* 258-262.
Lewis M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.
Lieberman MB, Asaba S (2006). Why do firms imitate each other? *Academy of Management Review*. 31(2) 366-385.
Mani D, Barua A, Whinston AB (2010). An empirical analysis of the impact of information capabilities design on business process outsourcing performance. *Management Information Systems Quarterly*. 34(1) 5.
March JG (1991). Exploration and exploitation in organizational learning. *Organization science*. 2(1) 71-87.
Martin R (2009). The design of business. *Harvard Business School Publishing, Massachusetts*.
McAfee A (2002). The impact of enterprise information technology adoption on operational performance: an empirical investigation. *Production and operations management*. 11(1) 33-53.
McElheran K, Brynjolfsson E (2015). Data in Action: Data-Driven Decision Making in US Manufacturing.
McElheran KS (2011). Do market leaders lead in business process innovation? The case (s) of e-business adoption. *Harvard Business Review*.
Melville N, Kraemer K, Gurbaxani V (2004). Review: Information technology and organizational performance: An integrative model of IT business value. *MIS quarterly*. 28(2) 283-322.
Milgrom P, Roberts J (1990). The economics of modern manufacturing: Technology, strategy, and organization. *The American Economic Review* 511-528.
Mims C (2015). *Data is the New Middle Manager*. Online.
Mithas S, RAMASUBBU N, Sambamurthy V (2011). How information management capability influences firm performance. *MIS quarterly*. 35(1) 237.
Moon Y. (2010). *Different: Escaping the competitive herd*. Crown Business.
Mucherino A, Papajorgji PJ, Pardalos PM. (2009). *Data mining in agriculture*. Springer Science & Business Media.
Murnane RJ, Levy F, Autor D (1999). *Technological change, computers and skill demands: Evidence from the back office operations of a large bank*.

Nonaka I, Von Krogh G (2009). Perspective-tacit knowledge and knowledge conversion: Controversy and advancement in organizational knowledge creation theory. *Organization science*. 20(3) 635-652.

Saunders A, Tambe P (2013). A Measure of Firms' Information Practices Based on Textual Analysis of 10-K Filings.

Silver N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Tambe P, Hitt LM (2013). Job hopping, information technology spillovers, and productivity growth. *Management Science*. 60(2) 338-355.

Tambe P, Hitt LM (2013). Measuring information technology spillovers. *Information Systems Research*. 25(1) 53-71.

Tambe P, Hitt LM, Brynjolfsson E (2012). The extroverted firm: How external information practices affect innovation and productivity. *Management Science*. 58(5) 843-859.

Wu L, Brynjolfsson E (2014). *The future of prediction: How Google searches foreshadow housing prices and sales*. University of Chicago Press.

Wu L, Jin F, Hitt LM (2014). Are All Spillovers Created Equal? A Network Perspective on IT Labor Movements.

Zuboff S. (1988). *In the age of the smart machine: The future of work and power*. Basic Books.